A shorter version of this tutorial appears as an appendix in Cubillos et al. (2015):

On Correlated-noise Analyses Applied to Exoplanet Light Curves
https://ui.adsabs.harvard.edu/abs/2017AJ....153....3C/abstract

Draft version December 4, 2015
Preprint typeset using LaTeX style emulateapj v. 8/13/10

NOTES ON CORRELATED NOISE AND DEPENDENT DATA RESAMPLING

Tom Loredo

1. THE IMPACT OF CORRELATED NOISE

A central ingredient in both frequentist and Bayesian parametric modeling is the sampling distribution for the data: the joint probability density function (PDF) for the data values, as a function of the model parameters. In our time series setting, we denote the data by $\vec{y} = (y_1, \ldots, y_n)$, with y_i denoting the value of a measurement at time t_i. Similarly, we denote the model predictions by $\vec{f} = (f_1, \ldots, f_n)$, with $f_i = f(t_i; \theta)$ for a model function with parameters θ. Note that the predictions are functions of the parameters, $f_i(\theta)$, but we often suppress the parameter dependence for convenience. The sampling distribution is the n-dimensional joint PDF, $p(\vec{y}|\theta)$.

When $\vec{y}$ is fixed to an actually observed data vector, the sampling distribution as a function of the model parameters is called the *likelihood function*, $\mathcal{L}(\theta)$. Bayesian methods quantify uncertainty in the parameters via the dependence of $\mathcal{L}(\theta)$ on the parameters; Bayes's theorem and the law of total probability convert this dependence into posterior probabilities for statements about the parameters. Frequentist methods quantify uncertainty by defining statistics (functions of $\vec{y}$) that produce point estimates or intervals in the parameter space (perhaps using the θ dependence of the likelihood function), and then using the sampling distribution to quantify the variability of the statistics across ensembles of hypothetical data vectors. The variability in the sample space then is mapped into uncertainty quantifications in the parameter space (e.g., bias of a point estimate, or coverage of a confidence interval).

Commonly, the data are modeled as the sum of the predictions and independent, zero-mean, normally-distributed noise,

$$y_i = f_i(\theta) + \epsilon_i, \tag{1}$$

or

$$\vec{y} = \vec{f}(\theta) + \vec{\epsilon}, \tag{2}$$

with independent noise probabilities

$$p(\epsilon_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{\epsilon_i^2}{2\sigma_i^2}\right], \tag{3}$$

where σ_i is the standard deviation for the noise contribution in measurement i. In this scenario, the sampling distribution factors,

$$\begin{aligned} p(\vec{y}|\theta) &= \prod_i p(y_i|\theta) \\ &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{\epsilon_i^2}{2\sigma_i^2}\right] \\ &\propto \left(\prod_i \sigma_i\right)^{-1} \exp\left[-\frac{Q(\theta)}{2}\right], \end{aligned} \tag{4}$$

where $Q(\theta)$ is the familiar "χ^2" sum of squared, standardized residuals,

$$Q(\theta) = \sum_i \frac{(y_i - f_i)^2}{\sigma_i^2}. \tag{5}$$

To expose the underlying geometry of model fitting in this setting, we introduce the (diagonal) covariance matrix for the noise, $\boldsymbol{C}$, with components

$$\begin{aligned} C_{ij} &\equiv \langle \epsilon_i \epsilon_j \rangle \\ &= \sigma_i \sigma_j \delta_{ij}, \end{aligned} \tag{6}$$

and its inverse, the *precision* or *concentration matrix* $\boldsymbol{K}$, with components

$$\begin{aligned} K_{ij} &\equiv \boldsymbol{C}^{-1} \\ &= \frac{1}{\sigma_i \sigma_j} \delta_{ij}. \end{aligned} \tag{7}$$

Now we can write

$$\begin{aligned} Q(\theta) &= (\vec{y} - \vec{f})^2 \\ &\equiv (\vec{y} - \vec{f}) \cdot (\vec{y} - \vec{f}), \end{aligned} \tag{8}$$

where the dot product between two vectors $\vec{a}$ and $\vec{b}$ is defined by

$$\vec{a} \cdot \vec{b} \equiv \sum_{i=1}^{n} \sum_{j=1}^{n} a_i K_{ij} b_j \tag{9}$$

$$= \sum_{i=1}^{n} \frac{a_i b_i}{\sigma_i^2}, \tag{10}$$

with the collapse to a single sum on the last line a consequence of independence. The precision matrix thus plays the role of a metric in sample space, defining how far apart two vectors of observed or predicted y_i values are.

1.1. *Linear models and least squares estimation*

We want to gain insight into how the precision matrix influences inferences; we will thus focus on simple, analytically tractable models. Consider *linear models*, that is, models with predictions that are linear with respect to the parameters. Take $\theta = (A_1, \ldots, A_m)$, a set of amplitude coefficients expressing $f(t;\theta)$ in terms of component models $g_\alpha(t)$ (note the use of Greek letters to index the m-dimensional model space):

$$f(t;\theta) = \sum_{\alpha=1}^{m} A_\alpha g_\alpha(t), \tag{11}$$

or, defining $\vec{g}_\alpha = (g_\alpha(t_1), \ldots, g_\alpha(t_n))$,

$$\vec{f}(\theta) = \sum_{\alpha=1}^{m} A_\alpha \vec{g}_\alpha. \tag{12}$$

Example basis models include polynomials, where $g_\alpha(t)$ would comprise a polynomial basis (e.g., monomials, or an orthogonal family), or a (finite) Fourier series of sine and cosine functions at harmonic frequencies. The m model vectors span a subspace of the n-dimensional sample space of dimension at most m; we will assume there is no redundancy among the component models so this subspace has dimension m.

Presuming the noise standard deviations are known, the maximum likelihood parameter values (the mode of the posterior PDF, if a flat prior is adopted) are the values that minimize $Q(\theta)$. Combining equation (8) and equation (12), we seek to minimize

$$\begin{aligned} Q(\theta) &= \vec{y} \cdot \vec{y} + \vec{f}(\theta) \cdot \vec{f}(\theta) - 2\vec{y} \cdot \vec{f}(\theta) \\ &= y^2 + \sum_{\alpha} \sum_{\beta} A_\alpha A_\beta \vec{g}_\alpha \cdot \vec{g}_\beta - 2 \sum_{\alpha} A_\alpha \vec{y} \cdot \vec{g}_\alpha. \end{aligned} \tag{13}$$

The values $\hat{A}_\alpha$ that minimize Q make $\partial Q / \partial A_\alpha$ vanish; this condition implies

$$\sum_{\beta} \hat{A}_\beta \vec{g}_\beta \cdot \vec{g}_\alpha = \vec{y} \cdot \vec{g}_\alpha. \tag{14}$$

Denoting the resulting best-fit function vector as $\hat{\vec{f}} = \sum_\beta \hat{A}_\beta \vec{g}_\beta$, this condition corresponds to

$$\hat{\vec{f}} \cdot \vec{g}_\alpha = \vec{y} \cdot \vec{g}_\alpha, \tag{15}$$

that is, *the best-fit model is the model whose projection on each basis function equals the data's projection on each basis function.* Geometrically, it is the projection of the data vector onto the model subspace, or equivalently, the point in the model subspace closest to the point in the full samples space corresponding to the data.

To solve explicitly for the amplitudes, we introduce the *model metric* matrix, $\boldsymbol{\eta}$, with components $\eta_{\alpha\beta} = \vec{g}_\alpha \cdot \vec{g}_\beta$ (this is an $m \times m$ matrix, in contrast to the $n \times n$ matrices, $\boldsymbol{C}$ and $\boldsymbol{K}$). Then equation (15) implies

$$\hat{A}_\alpha = \sum_{\beta} [\boldsymbol{\eta}^{-1}]_{\alpha\beta} \, \vec{y} \cdot \vec{g}_\beta. \tag{16}$$

This shows that *the best-fit amplitudes are built from weighted projections of the data onto the basis functions*; the weights are determined by the projections of the basis functions onto each other, via $\boldsymbol{\eta}$. Using the fact that $\boldsymbol{K}$ is diagonal, we can write this projection more explicitly as

$$\hat{A}_\alpha = \sum_{\beta} [\eta^{-1}]_{\alpha\beta} \sum_i \frac{y_i g_{\beta i}}{\sigma_i^2}. \tag{17}$$

When the noise is homoskedastic (so $\sigma_i = \sigma$), these estimates are called *ordinary least squares* (OLS) estimates; estimates in the heteroskedastic case are weighted OLS estimates.

A virtue of the geometric notation is that the generalization to correlated Gaussian noise is notationally trivial. Correlated Gaussian noise has a multivariate normal PDF with a non-diagonal precision matrix,

$$p(\vec{\epsilon}) = \frac{|\boldsymbol{K}|^{1/2}}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_{ij} \epsilon_i K_{ij} \epsilon_j\right], \tag{18}$$

where $|\boldsymbol{K}|$ denotes the determinant of the precision matrix, and the double sum no longer collapses to a single sum, due to $\boldsymbol{K}$ having nonzero off-diagonal terms when there is correlated noise. Because we have defined sample space vector dot products in terms of $\boldsymbol{K}$ (see eqn. (9)), none of the equations above that are expressed in vector notation need to change to account for correlated noise. However, *the dot product has been redefined*, changing the nature of the projections appearing above. Equations such as (10) and (17), expressed in terms of the vector components, and invoking independence to collapse double sums to single sums, are now more complicated. In particular, the latter is

$$\hat{A}_\alpha = \sum_\beta [\eta^{-1}]_{\alpha\beta} \sum_{ij} y_i K_{ij} g_{\beta j}. \tag{19}$$

Noise correlations alter the nature of the projections we must make to estimate parameters and quantify parameter uncertainty. Linear model estimates found using these altered projections are called *generalized least squares* (GLS) estimates.

Methods currently used to handle correlated noise in *Spitzer* photometry data fall into two broad classes. Methods like the CW09 wavelet approach estimate a correlation matrix (and thus a precision matrix), and produce GLS estimates. In contrast, the time averaging and residual permutation approaches rely on OLS for estimation, but devise rules for inflating uncertainties to account for correlations ignored by OLS. It has been demonstrated that OLS estimates are statistically consistent; that is, asymptotically (as $n \to \infty$), OLS estimates converge to the true parameter values. Roughly speaking, although correlation complicates the way information accumulates across samples, infinite sample size ameliorates the complications. However, because OLS does not use projections that account for correlations, the quality of estimates can be significantly compromised with finite sample size.

1.2. *The AR(1) hidden Markov model*

To gain insight into the difference between OLS and GLS estimates, we consider examples with a simple correlated noise model: AR(1) autoregressive noise, for regularly sampled data. Our treatment adapts analyses by Zellner (1971) and Sivia and Skilling (2006) on related models. In this model, the conditional expectation (regression) of the noise for sample i is proportional to the previous noise value; the actual value of the noise is the sum of this expectation and a new zero-mean *innovation* contribution:

$$\epsilon_i = \phi \epsilon_{i-1} + \nu_i, \tag{20}$$

where ϕ is the autoregression parameter, and ν_i is the innovation. The innovations are independent, with zero-mean normal PDFs with standard deviation s. The overall model for $f(\theta)$ is a *hidden Markov model* (HMM): "Markov" indicating that the prediction for the noise at time t_i depends only on the noise at the previous time, and not on the whole noise history; and "hidden" because ϵ_i is not directly observed (as it would be in a standard AR(1) model), rather, y_i is observed, mixing uncertain model and noise contributions. This HMM is a simple example of a *state space model* (SSM), a class of models comprised of two components: an evolution model for an unobserved system state (corresponding to eqn. (20) here), and an observation model linking the state to observables (corresponding to (1)).

The AR(1) model enables recursive construction of the joint distribution for the noise. The model specifies independent normal PDFs for the ν_i terms, so the goal is to express the ϵ_i values entirely in terms of ν_i values. The probability for the first noise sample, ϵ_1, is slightly complicated by the fact that it depends on innovations at times before there is data. However, ϵ_i is a linear sum of terms that are each zero-mean normal, so it must itself have a normal PDF, with variance given by the sum of the variances of its contributions. Writing $\epsilon_{i-1} = \nu_{i-1} + \phi_{i-1}$, and recursing, we find

$$\epsilon_1 = \sum_{j=0}^{\infty} \phi^j \nu_{1-j}. \tag{21}$$

The standard deviation of each term is $\phi^j s$, so the sum of the variances is

$$\begin{aligned} \sigma_\epsilon^2 &= s^2 \sum_{j=0}^{\infty} \phi^{2j} \\ &= \frac{s^2}{1-\phi^2}, \end{aligned} \tag{22}$$

provided that $|\phi| < 1$ (otherwise the series diverges). The marginal PDF for ϵ_i at any time is a zero-mean normal with this variance; the noise time series is thus *stationary* (with the same marginal distribution at each time). Note

that when $\phi = 1$, we have $\epsilon_i = \epsilon_{i-1} + \nu_i$, the definition of a Gaussian random walk, for which the standard deviation grows like $\sqrt{t}$. This is nonstationary behavior, and $|\phi| = 1$ marks the boundary between stationary and nonstationary AR(1) models.

We can always write the joint PDF for all noise values in terms of factors that condition on the previous history. Let $\epsilon_{i:j} = (\epsilon_i, \ldots, \epsilon_j)$; then we can write

$$p(\vec{\epsilon}) = p(\epsilon_1)\, p(\epsilon_2|\epsilon_1)\, p(\epsilon_3|\epsilon_{1:2}) \cdots p(\epsilon_n|\epsilon_{1:n-1}). \tag{23}$$

But the Markov property of the AR(1) noise model implies that this simplifies to

$$p(\vec{\epsilon}) = p(\epsilon_1) \prod_{i=2}^{n} p(\epsilon_i|\epsilon_{i-1}). \tag{24}$$

Equation (20) implies that $p(\epsilon_i|\epsilon_{i-1})$ is the probability that $\nu_i = \epsilon_i - \phi\epsilon_{i-1}$. The factors appearing in equation (24) are thus

$$p(\epsilon_1) = \frac{1-\phi^2}{s\sqrt{2\pi}} e^{-\epsilon_1^2/2s^2}; \tag{25}$$

$$p(\epsilon_i|\epsilon_{i-1}) = \frac{1}{s\sqrt{2\pi}} \exp\left[-\frac{1}{2s^2}(\epsilon_i - \phi\epsilon_{i-1})^2\right]. \tag{26}$$

The observation equation, equation (1), indicates that the probability for the data, $\vec{y}$, is the probability that the noise values take on the values $\epsilon_i = y_i - f_i(\theta)$. Let $r_i(\theta) \equiv y_i - f_i(\theta)$ denote the residuals from adopting the model with parameters θ. Then the PDF for the data can be written

$$p(\vec{y}|\theta) = \frac{(1-\phi^2)^{1/2}}{s^n(2\pi)^{n/2}} e^{-Q(\theta)/2s^2}, \tag{27}$$

with

$$\begin{aligned} Q(\theta) &= (1-\phi^2)r_1^2 + \sum_{i=2}^{n}(r_i - \phi r_{i-1})^2 \\ &= \sum_{i=1}^{n} r_i^2 + \phi^2 \sum_{i=2}^{n-1} r_i^2 - 2\phi \sum_{i=2}^{n} r_i r_{i-1}. \end{aligned} \tag{28}$$

The first term—the sum of squared residuals—is just the "χ^2" term that appears in OLS (see eqn. (5)). When $\phi \neq 0$, AR(1) noise correlations introduce new contributions to $Q(\theta)$, including a term resembling a lag-1 autocorrelation (the last term). These terms correspond to changes in the model basis projections entailed by the correlations in a GLS analysis, versus an OLS analysis.

Since the precision matrix element K_{ij} is just the coefficient of the $r_i r_j$ term in Q/s^2, we can read off the precision matrix for AR(1) noise from equation (28):

$$\boldsymbol{K} = \frac{1}{s^2}\begin{bmatrix} 1 & -\phi & 0 & 0 & \cdots \\ -\phi & 1+\phi^2 & -\phi & 0 & \cdots \\ 0 & -\phi & 1+\phi^2 & -\phi & \cdots \\ \vdots & \ddots & \ddots & \ddots & \cdots \\ \cdots & 0 & 0 & -\phi & 1 \end{bmatrix}. \tag{29}$$

This has a tridiagonal form. The covariance matrix can be found by inverting $\boldsymbol{K}$ (using known results on inversion of tridiagonal matrices). More simply, note that an equation like equation (21) holds for any ϵ_i, expressing it in terms of the innovations:

$$\epsilon_i = \sum_{k=0}^{\infty} \phi^k \nu_{i-k}. \tag{30}$$

Exploiting the independence of the innovations, the covariance between ϵ_i and ϵ_j can be computed from the sum of the coefficients of squared innovations in the product $\epsilon_i\epsilon_j$ (since terms proportional to $\nu_k\nu_l$ will have zero covariance if $k \neq l$); this gives

$$C_{ij} = \frac{s^2\phi^{|i-j|}}{1-\phi^2}. \tag{31}$$

This shows that the magnitude of the correlation between noise values falls exponentially with separation in time (in the stationary regime, where $|\phi| < 1$). For $\phi < 0$, the correlation flips sign with each increase in lag, giving rise to somewhat oscillatory behavior in the noise.

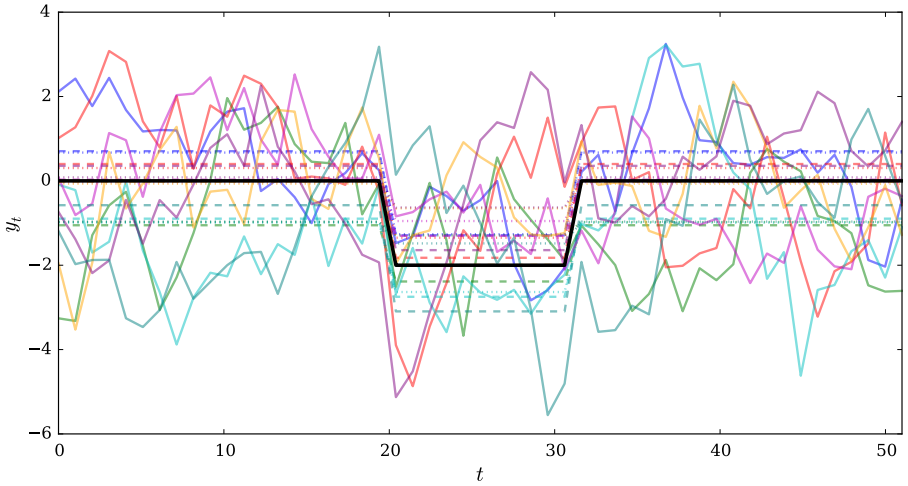

FIG. 1.— Eight sampled time series from the baseline-dip model with AR(1) noise. Curves of matching color connect the simulated data (solid), the GLS best-fit points (dashed), and the OLS best-fit points (dotted). The solid black curve shows the true (noiseless) function.

1.3. *Example: Constant signal*

The simple case of a constant signal model of unknown amplitude, $f(t;\mu) = \mu$, is analytically tractable and is illuminating. Substituting $r_i = y_i - \mu$ and minimizing $Q(\mu)$ leads to the GLS estimate

$$\hat{\mu} = \frac{w_n \bar{y} + w_2 (y_1 + y_n)/2}{w_n + w_2}, \tag{32}$$

where $\bar{y}$ is the sample mean, $\bar{y} \equiv (1/n)\sum_i y_i$, and we have defined weights $w_n = n(1-\phi)$ and $w_2 = 2\phi$. When $\phi = 0$ (independent noise), $\hat{\mu}$ is just the sample mean. Otherwise, $\hat{\mu}$ is a weighted average of the full sample mean, and the average of the first and last (i.e., the most widely separated and thus least correlated) samples. As ϕ approaches unity (strongly positively correlated noise), GLS instructs us to just average the most widely separated samples. In contrast, the OLS estimate is always the full sample mean. The OLS and GLS estimates thus will differ, not just in the uncertainties they assign to the mean, but also in the actual values of the estimates.

$Q(\mu)$ is quadratic in μ, so the likelihood function is a Gaussian function in μ. The reciprocal of the second derivative of $Q(\mu)/s^2$ at $\hat{\mu}$ gives the squared standard deviation of this Gaussian,

$$\sigma_\mu^2 = \frac{s^2}{n(1-\phi)^2 - 2\phi(1-\phi)}. \tag{33}$$

When $\phi = 0$, we have $\sigma_\mu = s/\sqrt{n}$, the familiar "root-n" result. As ϕ approaches unity, the denominator decreases toward zero, and the uncertainty in μ grows.[1] Roughly speaking, growing positive correlation decreases the effective sample size, inflating uncertainties. This motivates approaches like time averaging that attempt to account for correlation merely by inflating uncertainties. But such approaches do not account for the effect of correlations on the actual value of a finite-sample estimate.

1.4. *Example: Constant baseline with dip*

The effect of correlations on parameter estimates depends on the extent to which the correlations may mimic or distort the projections of the data onto the model components. When a model has components that vary slowly with respect to the correlation scale, the main effect of correlations is to change the effective sample size. But when a model has temporally localized components, correlations can significantly affect, not just the uncertainty scale, but also the best-fit parameter values.

To illustrate this, we used simulated AR(1) noise and the GLS likelihood function to model data generated from a baseline signal of amplitude a, with a localized dip of depth δ. We took the dip location and width to be known. For the illustration we report here, we simulated 51 observations with true parameter values $\theta = (a, \delta) = (0, 2)$, with the dip spanning 10 samples in the middle of the time series. The noise was generated with an innovation standard deviation $s = 1$, and $\phi = 0.8$, producing data with autocorrelation time scales ~ 5. Figure 1 displays examples of the simulated data and OLS and GLS best-fit function estimates. It is visually apparent that the OLS and GLS estimates sometimes differ.

Figure 2 shows contours of the posterior PDFs for (a, δ) from two representative simulations. The left panel shows a case where the OLS and GLS best-fit estimates did not differ too dramatically. Even in this case, it is apparent that the OLS likelihood function not only has an incorrect uncertainty scale (which one might hope to fix via inflation),

[1] Note that there are different ways to take this limit. We can fix the innovation scale, s, and let $\phi \to 1$. This implies that the marginal noise variance, σ_ϵ^2 in equation (22), diverges, which contributes to inflation of the μ uncertainty. Alternatively, we can fix σ_ϵ, in which case s decreases toward zero as $\phi \to 1$.

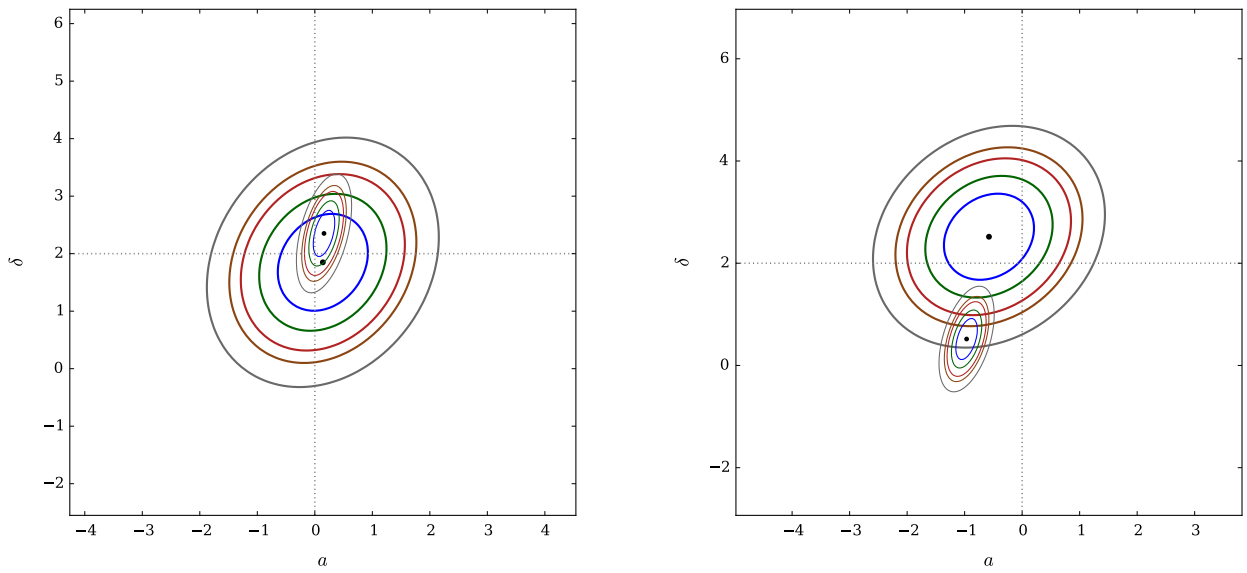

FIG. 2.— Contours of the posterior PDF for (a, δ), from GLS (larger, thicker contours) and OLS (smaller, thinner contours) analyses for two representative simulations. From inside to outside, the contours bound highest posterior density (HPD) regions with 50%, 75%, 90%, 95%, and 99% of the posterior probability. Dots indicate the modes. Crosshairs indicate the true parameter values.

but does not correctly capture the shape of the PDF (i.e., the correlation between a and δ estimates). The right panel shows that the OLS and GLS estimates sometimes dramatically differ; here the OLS best-fit estimate is just outside of the 98% GLS credible region.

To quantify how well or poorly the estimates behave, we simulated 5000 datasets and compared the resulting OLS and GLS estimates. Figure 3 shows the best-fit parameter estimates and offsets for the first 200 simulations. It is apparent that the OLS and GLS estimates can differ significantly, and that the OLS estimates tend to be further from the true values than the GLS estimates. To quantify the behavior of the estimates, we calculated two "credibility distances" based on the posterior PDFs for each case. For the GLS estimates, we calculated the probability C for a credible region that just includes the true parameter values. Accurate models should produce approximately calibrated inferences (i.e., regions with coverage close to the credible level), in which case the C values should be uniformly distributed among the simulations. The green histogram in Figure 4 (left panel) shows that the GLS estimates are in fact well-calibrated. For the OLS estimates, we calculated the probability C for the GLS credible region that just includes the OLS best-fit point; this measures the distance of the OLS inference from the well-calibrated GLS inference. If the OLS estimates matched the GLS estimates, these C values would all be near zero. Instead, the red histogram shows that the OLS estimates are often far from the GLS estimates. The center panel shows the cumulative distribution of the OLS C distances, indicating how often the OLS estimate is greater than any C value of interest. For example, it shows that the OLS best-fit estimate is outside the GLS 50% credible region about 40% of the time, and it is outside the GLS 90% region about 11% of the time. The right panel shows results from an additional simulation with a somewhat stronger correlation, $\phi = 0.9$. The OLS estimates become significantly worse, with the OLS best-fit estimate more often lying outside a large GLS credible region (e.g., the OLS estimates now lie outside the GLS 90% region about 25% of the time).

Finally, Figure 5 shows the frequency distributions of the errors in the best-fit estimates of the dip depth, δ, using the two methods. As was apparent from the scatterplot in Figure 3, neglecting correlation shifts the OLS estimates away from the truth, leading to significantly larger errors on average. In this case the OLS root-mean-square error (RMSE) is 68% larger than the GLS RMSE.

The main message of these examples is that noise correlation not only can inflate uncertainties; it can also corrupt parameter estimates, particularly when parameters of interest pertain to temporally localized structure in the model, for which noise correlations can significantly change the data projections needed for accurate inference. Methods that seek to account for correlations only by inflating parameter uncertainties are at best suboptimal (producing larger estimation errors than could be achieved with a good correlated noise model), and can sometimes be significantly misleading.

The analyses above considered the noise correlation structure to be known—both that it was AR(1), and what values the autocorrelation parameter, ϕ, and the innovation standard deviation, s, took. When the correlation structure is unknown, additional uncertainties arise from having to infer it. In a correlated Gaussian noise setting, the fundamental task is inferring the covariance matrix (or precision matrix) for the noise. For a time series of n samples, the (symmetric) covariance matrix has $\approx n^2/2$ entries; it is not possible to infer these from n samples without strong knowledge or assumptions about the covariance structure. The wavelet-based approach of CW09 parameterizes the covariance matrix by means of a wavelet decomposition into components of varied scale and location, with a parametric distribution

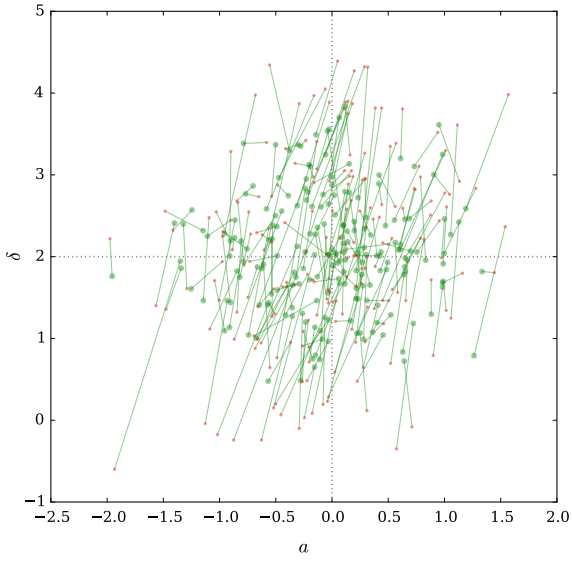

FIG. 3.— Best-fit OLS and GLS parameter estimates for 200 simulations. Green dots show the GLS estimates, connected by green lines to OLS estimates shown as small red dots.

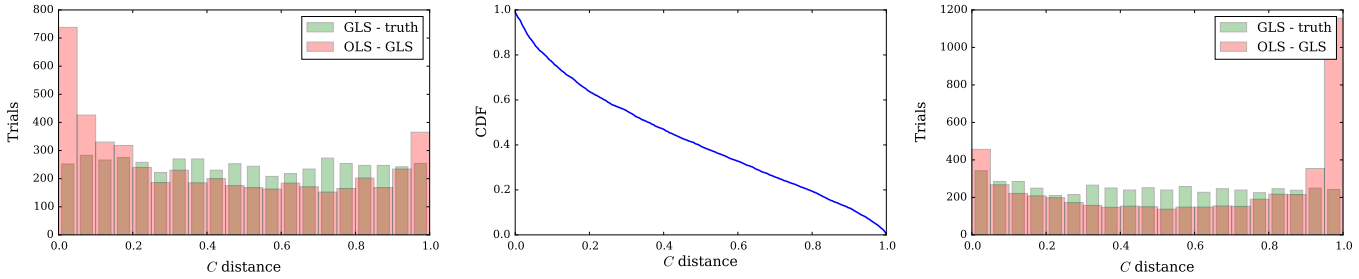

FIG. 4.— Distributions of credibility distance measures of quality of GLS and OLS inferences. Left and middle panels are from simulations with $\phi = 0.8$. The left panel green histogram shows the frequency distribution for the probability C of GLS credible regions that just include the true parameter values (plotted bins are slightly reduced in width for visibility). The red histogram shows the frequency distribution for the probability C of GLS credible regions that just include the OLS best-fit parameter values; this measures how far the OLS estimate is from the more accurate GLS estimate. Center panel shows the cumulative distribution of the OLS C values. The right panel duplicates the left panel, but for simulations with $\phi = 0.9$.

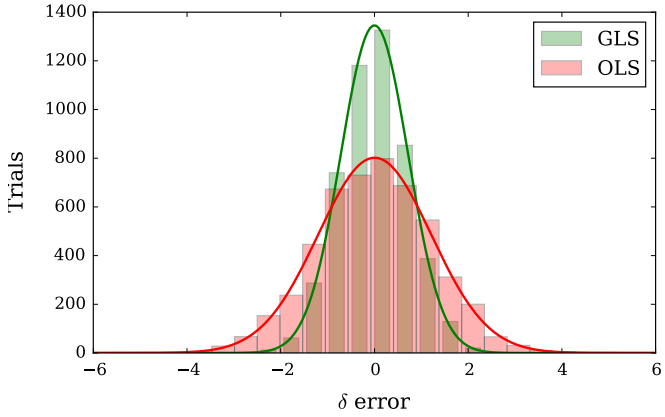

FIG. 5.— Frequency distributions for error in the dip depth estimates, for GLS estimates (green histogram, with bins narrowed for visibility), and for OLS estimates (red histogram). Solid curves show normal distribution fits to the frequency distributions.

imposed on the PDF for the amplitudes of those components. Other approaches have been developed for data from ground-based observations of transits and eclipses, including autoregressive moving average (ARMA) models that operate much like the AR(1) model described above [REF-Jordan], and Gaussian process (GP) approaches that assume stationarity and a few-parameter model for the covariance matrix components [REF-Gibson]. Here we focus on the CW09 approach. . . [say why—has it been used for Spitzer data previously?].

2. DEPENDENT DATA RESAMPLING

In frequentist statistics, parameter uncertainty is quantified by exploring the variability of point and interval estimators applied to datasets independently drawn from from the sampling distribution, $p(\vec{y}|\theta)$. By "independently drawn," we mean that entire *vectors* $\vec{y}$ are drawn from the sampling distribution independently from each other. To avoid confusion between sampling in time (with each y_i called a sample), and sampling an entire observation vector (i.e., draw a vector from $p(\vec{y}|\theta)$), we refer to each draw from $p(\vec{y}|\theta)$ as a *replication*—a hypothetical repetition of the entire time series observation. Similarly, in this section we refer to the sampling distribution, $p(\vec{y}|\theta)$, as the *replication distribution*.

In simple contexts with independent and identically distributed (IID) noise contributions in each sample, the n-dimensional replication distribution may be written as the product of independent factors,

$$p(\vec{y}|\theta) = \prod_{i=1}^{n} f(y_i|\theta), \tag{34}$$

where $f(y|\theta)$ is the replication distribution for single datum. In such settings, a single replication may be generated by making n independent draws of y_i from $f(y|\theta)$. If the noise distributions are independent but heteroskedastic, a similar factorization may hold for standardized data (with y_i scaled by σ_i). But when measurements include correlated noise, this factorization is not possible.

We would like to study variability using replications corresponding to θ near the true value, θ^*, but of course we do not know θ^*. Bootstrap methods use the data to construct a replication distribution that is close to $p(\vec{y}|\theta^*)$. *Parametric bootstrapping* (PB) applies when a sound parametric replication distribution is available, i.e., a PDF $p(\vec{y}|\theta)$. PB uses a parameter estimate based on the observed data, $\hat{\theta}(\vec{y}_{\rm obs})$, as a surrogate for θ^*; PB theory shows how to use replications from $p(\vec{y}|\hat{\theta}(\vec{y}_{\rm obs}))$ to compute good approximate confidence intervals and unbiased estimators. *Nonparametric bootstrapping* (NPB) applies when a sound parametric replication PDF is *not* available. NPB methods directly use the observed sample values, $\vec{y}_{\rm obs}$, to construct a replication distribution that approximates $p(\vec{y}|\theta^*)$. Sampling from this distribution typically involves *resampling*—generating hypothetical replications by random reuse of the observed samples.

In the IID sample setting, NPB methods typically rely on independent resampling of the data or residuals (possibly scaled), with replacement (see Thomas et al. 2016 for an astronomical example using scaled, resampled residuals). This corresponds to approximating the factors in the replication distribution by a sum of δ functions using the observed values. For example, when bootstrapping the data values directly, the bootstrap replication distribution is

$$p_b(\vec{y}) \equiv \prod_{i=1}^{n} f_b(y_i), \tag{35}$$

with the single-sample bootstrap PDF given by

$$f_b(y) = \frac{1}{n}\sum_j \delta(y - y_{{\rm obs},j}), \tag{36}$$

which is intended to approximate $p(y|\theta^*)$. Thus a $\vec{y}$ replication is created by drawing each y_i value from the same bootstrap sample PDF, with the draw made simply by choosing one of the observed data values at random. This approach obviously is not applicable when there is correlated noise, as it assumes independence and takes no account of the time ordering of the data.

Note that resampling *with replacement* is crucial to making NPB work; it makes the replications independent draws from $p_b(\vec{y})$, which bootstrap theory shows mimics the variability that would be seen among independent draws from the true replication distribution. Were the resampling done without replacement, each replication would have exactly the same set of data values (albeit permuted), and for problems with IID noise, there would be no variability at all in the ensemble of estimates or intervals created by resampling without replacement.

The prayer bead method is evidently motivated by the NPB idea of using the observed data to generate replications[[; it may also be motivated by permutation tests, a family of statistical methods related to the bootstrap]]. [[**Probably remove the permutation remark if we don't elaborate on it; see below.**]] The motivating idea is that shifting the data preserves time ordering and thus correlation structure. While this is true, the shifted datasets do not correspond to independent replications from any distribution, and thus do not exhibit the variability necessary for uncertainty quantification (e.g., computing confidence levels or estimator bias). For example, if the signal being estimated is simply a constant (as in the AR(1) HMM example above; see eqn. (32)), the prayer bead method would produce no variability at all in the parameter estimates. The prayer bead method is unsound as a tool for quantifying

uncertainty in parameter estimates, because it does not produce ensembles that mimic the behavior of independent $\vec{y}$ draws from a replication distribution.

There is a significant literature on generalizing the IID nonparametric bootstrap idea to address time series problems with correlated noise; this is a topic of ongoing research. One widely used approach is the *block bootstrap*. Presuming the investigator knows or can estimate a longest scale for correlations, Δt, the data are divided into blocks of length greater than Δt, and bootstrap resampling is done by drawing *blocks* at random (with replacement) to build a replication. A particular block rigidly preserves the time ordering of a subset of the data; in replications, it will appear shifted in time by various amounts. In this respect block bootstrapping resembles the prayer bead method. But block resampling produces greater variability than shifting the entire data vector, and by sampling with replacement, it produces ensembles that approximate independent draws from a (dependent) replication distribution. The "segmented bootstrap" devised by Jenkins, Caldwell, and Borucki (2002) for analysis of ground-based transit photometry is similar to the block bootstrap. The block bootstrap only works if the correlation scale is significantly shorter than the span of the data, which will often not be true for *Spitzer* exoplanet eclipse data, so we do not consider it further here. Further details about the block bootstrap and other methods for resampling dependent data may be found in Lahiri (2003).

[[I can add comments here about permutation tests, but the prayer bead method isn't much like a permutation test. The shifted datasets do correspond to permutations of the data, but only to a small subset of all possible permutations. Real permutation tests use all permutations (or a representative random sample of them); they also address certain specific types of hypothesis testing problems, not parameter uncertainty quantification. Let me know if you'd like this fleshed out.]]